A glimpse at the organization of the protein universe

Michele Vendruscolo* and Christopher M. Dobson*
Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

The amino acid sequences of most natural proteins result in an ability to fold to specific structures that generate biological activity, and simultaneously to avoid misfolding and aggregation (1). It appears from the data available to us at present that the overall architecture (the "fold") of these structures is much more highly conserved during evolution than the sequences that encode them. These folds have therefore emerged as ideal candidates for classifying proteins (Fig. 1) and hence to begin to make order of the protein universe (2). The continuing advances in structural biology, and particularly the recent emergence of structural genomics initiatives in which particular emphasis is placed on the discovery of new folds (3), are providing an opportunity to build up a comprehensive map of the protein universe. Of particular significance is the fact that the number of distinct structural archetypes, or folds, is thought to be relatively small, less than $\approx$10,000 by most estimates, with many different sequences able to encode the same basic fold of the polypeptide chain (4). A key question in the analysis of protein sequences and structures is the way in which they relate to their functions. Clues as to the answer will not only begin to enlighten us as to the fundamental organization of the protein universe, and the location within it of natural proteins, but will also provide a means of predicting the functions of those proteins for which this information is not yet defined by experiment. The ability to predict function will be of tremendous value, for example, in interpreting the output of genome sequencing programs, or in the design of new proteins with specific functional characteristics. In a recent issue of PNAS, Kim and colleagues (5) take a significant step toward this objective by extending their earlier study (6) to show that proteins with similar functions can be found close together in the protein universe—provided that the latter is organized through structural considerations in a suitable way, termed the structure space map (SSM).

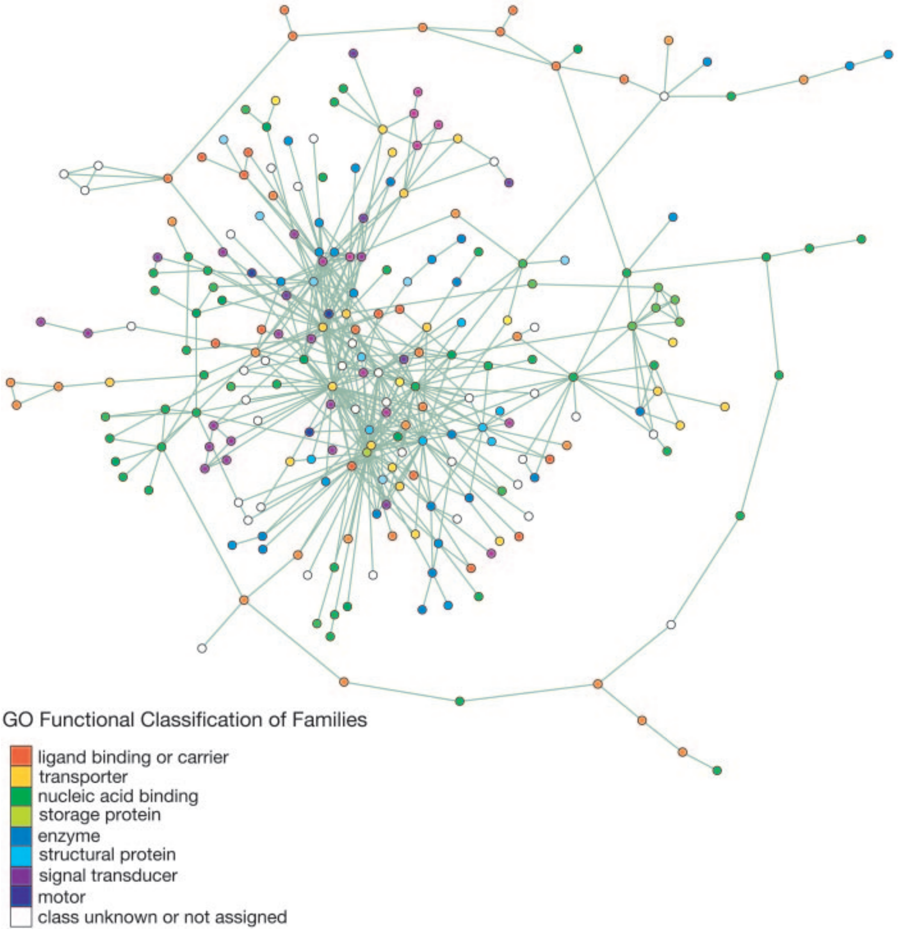

Fig. 1. The totality of all possible proteins can be looked at in three different ways: (*i*) the protein universe, which is formed by all possible amino acid sequences; (*ii*) the protein fold universe, which contains all possible folds associated with these sequences; and (*iii*) the protein function universe, which specifies all possible protein functions associated with these folds. The size of the protein universe is reduced from $\approx 10^{400}$ to $\approx 10^{10}$ different sequences if only those proteins currently in the biosphere are considered (4). In contrast, the protein fold universe may only contain some 10^5 different folds in total (23–25), and it may be that natural proteins already populate a very substantial fraction of them. The Gene Ontology (GO) Consortium lists $\approx 10^4$ possible types of functions (9), a number that can be taken as an approximate estimate of the size of the known part of the protein function universe. It is interesting to note that the number of known protein functions, which is a tiny fraction of the number of currently existing different protein sequences, is of the same order of magnitude of that of known protein folds and also of that of small molecules found in living organisms (15). As demonstrated by the work of Kim and colleagues (5), the use of structural or functional considerations suggests a similar organization of the protein universe. The figure illustrates how an organization of the protein universe based on the GO classification appears to be compatible with a structural classification (13). [Reproduced with permission from ref. 13 (Copyright 2003).]

Colocalization of Structure and Function in the Protein Universe

The results of Kim and colleagues show that when the protein universe is ordered according to the SSM method (5), a functional classification emerges as a by-product of this exercise, a finding that suggests that a common chart, capable of reporting both structure and function simultaneously, may exist for the protein universe. This remarkable conclusion is prompted by the observation that proteins with similar functions colocalize in the SSM. The key to the success of the method is a multidimensional scaling procedure that enables the calculation of the "distance" between protein structures by the selective use of information closely related to their function (5). This procedure is thus capable

See companion article on page 3651 in issue 10 of volume 102.

*To whom correspondence may be addressed. E-mail: mv245@cam.ac.uk or cmd44@cam.ac.uk.

of making effective use of the observation that a primary determinant of functionality is the presence of local motifs, as in the case of EF hands (for calcium binding) or catalytic triads (for proteolytic reactions) (7, 8).

As a consequence of the findings of Kim and coworkers (5) one might expect a similar ordering scheme to be derived when protein functions, rather than structures, are considered. Indeed, the Gene Ontology (GO) Consortium aims at organizing the universe of natural proteins through a functional classification (9). The existence of proteins that perform similar functions and yet have different structures (e.g., chymotrypsin and subtilisin), and of proteins that perform different functions depending on the cellular context (10), complicates this already challenging exercise by creating shortcuts in the protein universe that can potentially disrupt its order by linking together distant regions of space in the organizational structure established by this approach. The multidimensional scaling introduced by Kim and coworkers (5) should, however, help to overcome such problems, because it appears to minimize the relevance of these shortcuts by clustering structures according to their functionally relevant motifs. Moreover, it has been shown recently by using a range of different methods that the structural and the GO classification are to a large extent compatible (11–13).

The SSM approach should also enable the rapid classification of new protein folds identified through structural genomics initiatives as well as more conventional structural biology procedures, and in the process also may help the simultaneous classification of proteins in terms of function. This type of chart should, in addition, aid in the development of rational strategies for the discovery of new drugs. Most drugs act by interfering with a particular protein function, but a drug targeted toward one protein may also affect other proteins with similar recognition sites (14, 15). The SSM method might well help to avoid unwanted adverse effects on other proteins by simplifying the search for functionally relevant protein similarities. Given the complex and crowded molecular environment within living systems (16), such an achievement would be a great leap forward in the quest to carry out the rational design of therapeutic molecules.

We shall soon understand many of the features about the organization of the protein universe.

Expanding the Boundaries of the Protein Universe

The results of approaches such as those described by Kim and coworkers (5), along with recent ideas about the underlying reasons that the number of different folds appears remarkably small (17) and about the way that the information needed to fold proteins and to minimize the risks of misfolding (1, 18) is encoded in the sequence (19–22), suggest that we shall soon understand many of the most important features about the physical basis of the organization of the protein universe. In addition, just as natural molecular evolution is constantly probing the protein universe, we may also endeavor to search for proteins with new functionalities to carry out a wider range of chemical reactions. The results presented by Kim and coworkers (5) are a very significant example of how the confluence of experimental methods, including structural genomics and proteomics, and theoretical methods, including sequence and structure classification, in conjunction with ideas about biological evolution, can provide a framework of general principles that acts as a guide for exploration of the protein universe and extend significantly the arsenal of functions that proteins are able to perform.

1. Dobson, C. M. (2003) *Nature* **426,** 884–890.
2. Holm, L. & Sander, C. (1996) *Science* **273,** 595–602.
3. Vitkup, D., Melamud, E., Moult, J. & Sander, C. (2001) *Nat. Struct. Biol.* **8,** 559–566.
4. Koonin, E. V., Wolf, Y. I. & Karev, G. P. (2002) *Nature* **420,** 218–223.
5. Hou, J., Jun, S.-R., Zhang, C. & Kim, S.-H. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 3651–3656.
6. Hou, J. T., Sims, G. E., Zhang, C. & Kim, S. H. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 2386–2390.
7. Irving, J. A., Whisstock, J. C. & Lesk, A. M. (2001) *Proteins* **42,** 378–382.
8. Kasuya, A. & Thornton, J. M. (1999) *J. Mol. Biol.* **286,** 1673–1691.
9. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000) *Nat. Genet.* **25,** 25–29.
10. Copley, S. D. (2003) *Curr. Opin. Chem. Biol.* **7,** 265–272.
11. Shakhnovich, B. E., Dokholyan, N. V., DeLisi, C. & Shakhnovich, E. I. (2003) *J. Mol. Biol.* **326,** 1–9.
12. Pazos, F. & Sternberg, M. J. E. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 14754–14759.
13. Ouzounis, C. A., Coulson, R. M. R., Enright, A. J., Kunin, V. & Pereira-Leal, J. B. (2003) *Nat. Rev. Genet.* **4,** 508–519.
14. Buchanan, S. G. (2002) *Curr. Opin. Drug Discov. Dev.* **5,** 367–381.
15. Dobson, C. M. (2004) *Nature* **432,** 824–828.
16. Ellis, R. J. & Minton, A. P. (2003) *Nature* **425,** 27–28.
17. Hoang, T. X., Trovato, A., Seno, F., Banavar, J. R. & Maritan, A. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 7960–7964.
18. Vendruscolo, M. & Dobson, C. M. (2005) *Philos. Trans. R. Soc. London Ser. A* **363,** 433–450.
19. Vendruscolo, M., Paci, E., Dobson, C. M. & Karplus, M. (2001) *Nature* **409,** 641–645.
20. Ptitsyn, O. B. (1998) *J. Mol. Biol.* **278,** 655–666.
21. Lindorff-Larsen, K., Rogen, P., Paci, E., Vendruscolo, M. & Dobson, C. M. (2005) *Trends Biochem. Sci.* **30,** 13–19.
22. Bashford, D., Chothia, C. & Lesk, A. M. (1987) *J. Mol. Biol.* **196,** 199–216.
23. Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N. & Orengo, C. A. (2000) *Nat. Struct. Biol.* **7,** 991–994.
24. Gerstein, M. & Levitt, M. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 11911–11916.
25. Coulson, A. F. W. & Moult, J. (2002) *Proteins* **46,** 61–71.